Data Analysis Task Report

Problem Statement

As part of our sustainability efforts, we track the amount of greenhouse gas emissions (GHG) produced by our apparel factories. This can be influenced by the amount of production that occurs so we have collected data on both GHG and key production metrics.

If our target is to reduce total GHG by 10% by 2026, what recommendations do you have for how this can be achieved?

Data

The data consists of the following columns:

- **Factory**: ID to identify which of six factories the data relates to
- **Month**: data is recorded monthly
- **GHG**: greenhouse gas emissions
  - These could be from electricity (including to power sewing machines), vehicles, boiler, etc.
- **SAH**: standard allowed hours
  - The time a skilled worker should take to complete the produced garments under standard conditions, and is a key measure of business activity
  - More pieces produced = more time required for production = higher SAH
- **EFF**: efficiency
  - The ratio of expected time for production to the actual time taken
  - For example, a garment with 8 SAH that takes 10 hours to produce will count as 80% efficiency
- **WH**: total working hours
  - Includes employees both directly and indirectly involved in production
- **W**: number of workers

Limitations:

- Lack of data given. Features like (Age, Gender, Production) would help in better analysis.

Data Preprocessing:

- The data contains 396 rows and 7 columns
- There are no missing values present in the data
- There are 5 unique factories starting from Aug 2018 and an addition of 6th factory (1966) from Aug 2021.
- The data ranges from Aug 2018 to July 2024

Exploratory Data Analysis

1. Summarizing the Data by Factory

First we summarize the data according to the factory to get better insights.

	Factory	GHG_mean	GHG_total	SAH_mean	SAH_total	EFF_mean	WH_mean	WH_total	W_mean
2	761	17553.758333	1263870.6	152007.777778	10944560.0	0.695278	4.408381e+05	31740340.0	2007.152778
1	434	16557.852778	1192165.4	278502.430556	20052175.0	0.496250	1.137897e+06	81928580.0	5164.958333
0	150	9670.350000	696265.2	278988.694444	20087186.0	0.713472	7.946102e+05	57211935.0	3607.750000
3	1122	7395.201389	532454.5	147467.375000	10617651.0	0.496944	6.019893e+05	43343232.0	2732.819444
4	1396	6251.515278	450109.1	149878.708333	10791267.0	0.358889	8.598803e+05	61911380.0	3907.208333
5	1966	2990.391667	107654.1	24641.333333	887088.0	0.344444	1.433712e+05	5161362.0	652.611111

Factory 761:

- Highest total GHG emissions: 1,263,870.6 units.
- Average GHG emissions: 17,553.8 units.
- Moderate efficiency that is 0.695.
- Relatively low total SAH and WH compared to other high-GHG factories.

Factory 434:

- Second highest total GHG emissions: 1,192,165.4 units.
- Lowest efficiency that is 0.496.
- Highest SAH and WH total values, indicating high production but inefficient operations.

Factory 150:

- Moderate GHG emissions: 696,265.2 units.
- Highest efficiency that is 0.713.
- High total SAH and WH, indicating efficient high production.

Factories 1122, 1396, and 1966:

- Lower GHG emissions, especially Factory 1966 with significantly lower GHG (107,654.1 units).
- Lower efficiency in some cases (e.g., Factory 1966 with an average efficiency of 0.344).
- Smaller production scale, except Factory 1122 which has moderate SAH and WH totals.

Key Observations:

- **Efficiency:** There is a noticeable difference in efficiency across factories. Factories with higher GHG emissions tend to have lower efficiency.
- **Production Volume:** Higher SAH and WH generally correlate with higher GHG emissions, as seen in Factory 434.
- **Potential Levers:** Improving efficiency, particularly in Factory 434 and others with low efficiency, could be a key lever for reducing GHG emissions.

2. Outlier Removal

We check for the outliers in the dataset. There were 2 outliers present in GHG, WH and W columns respectively. We removed the outliers using Inter-Quartile Range Method. These outliers don't affect the data and there was a very negligent change in the mean of data.

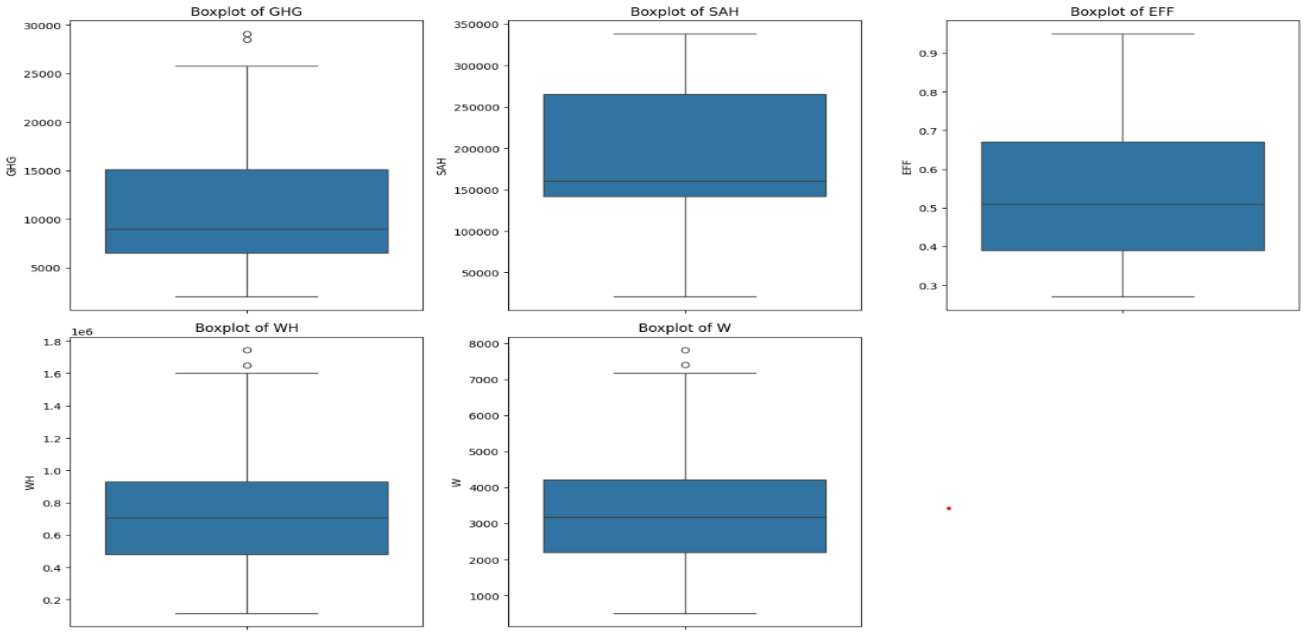


Figure 1 Data With Outliers

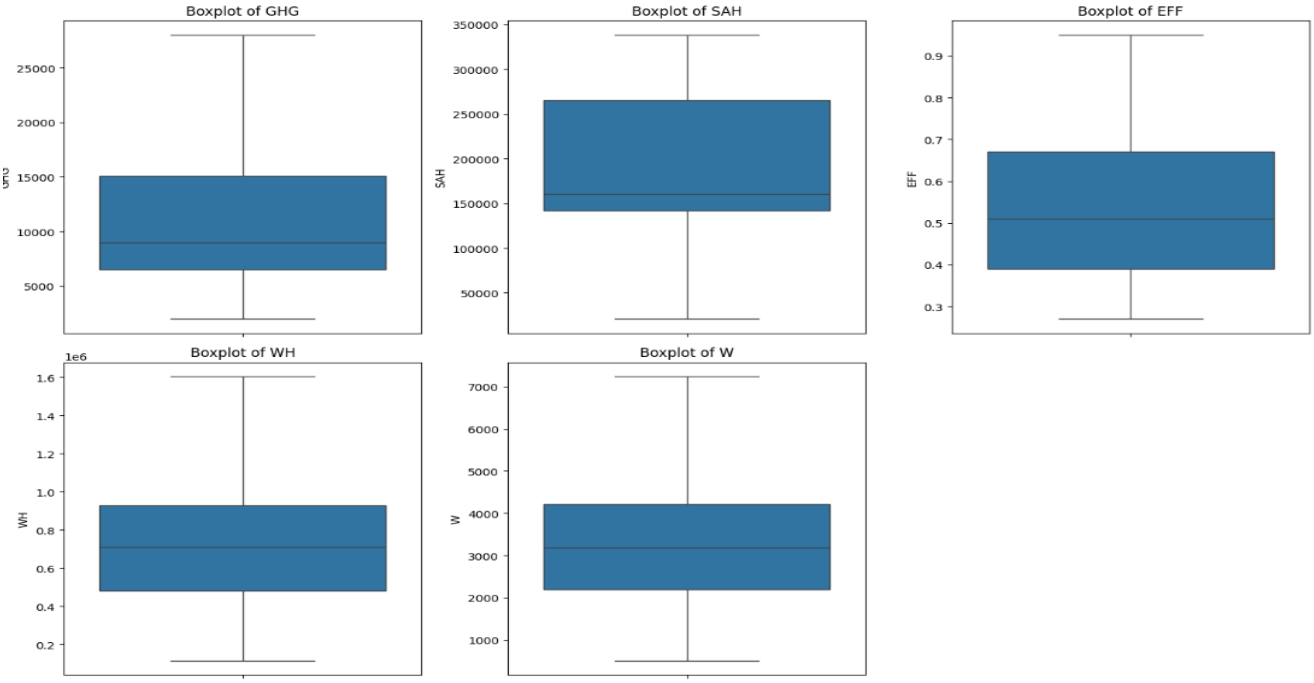


Figure 2 Cleaned Data With No Outliers

3. Average Worker, Standard Allowed Hours and Efficiency by Factory

Here we have analyzed how factories differ on the basis of their Worker, Standard Allowed hours and Efficiency. The box plot is created to visualize the results.

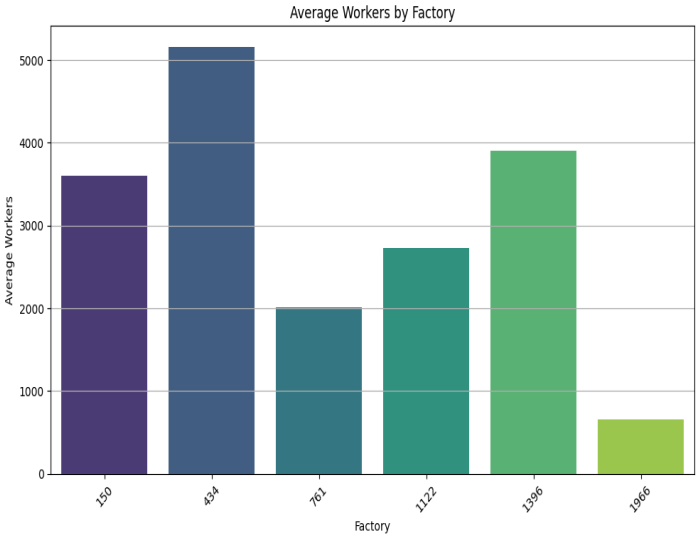


Figure 5 Average Worker By Factory

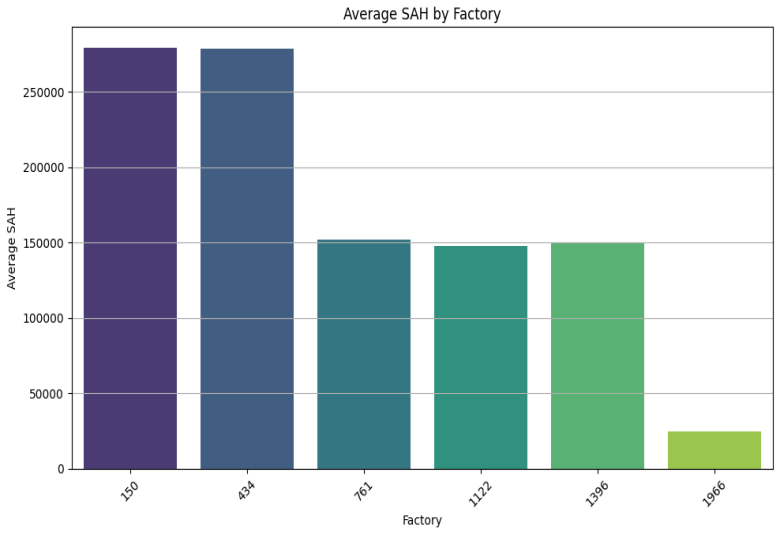


Figure 4 Average SAH By Factory

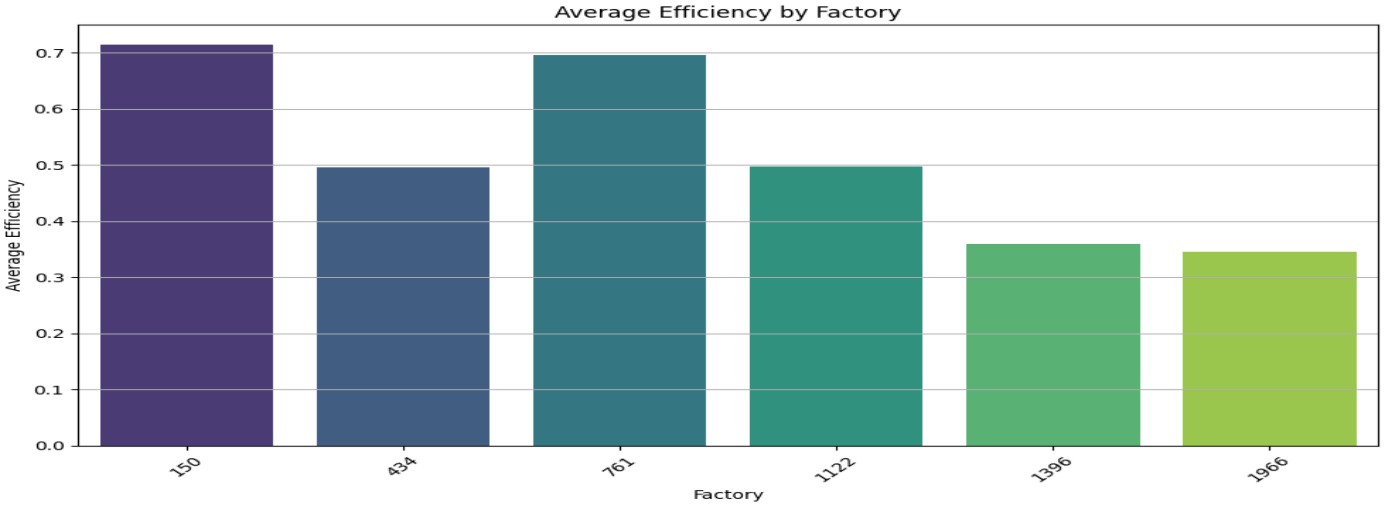


Figure 3 Average Efficiency By Factory

Key Observations:

- **Factory 434** has the highest no. of workers whereas **Factory 761** has the second lowest no. of workers
- **Factory 150** and **434** have the highest standard allowed hours as compared to other factories.
- **Factory 150** has the highest efficiency
- **Factory 761** has the second highest efficiency although having the second lowest no. of workers and moderate standard allowed hours.
- **Factory 434** having the highest no. of workers and standard allowed hours have the moderate efficiency.

4. Trend Analysis

We want to see the GHG emission over the years with resect to Factory and see if there is any specific trend that follows. For this, we created line plot, different lines showing the GHG emission with respect to factories over the year

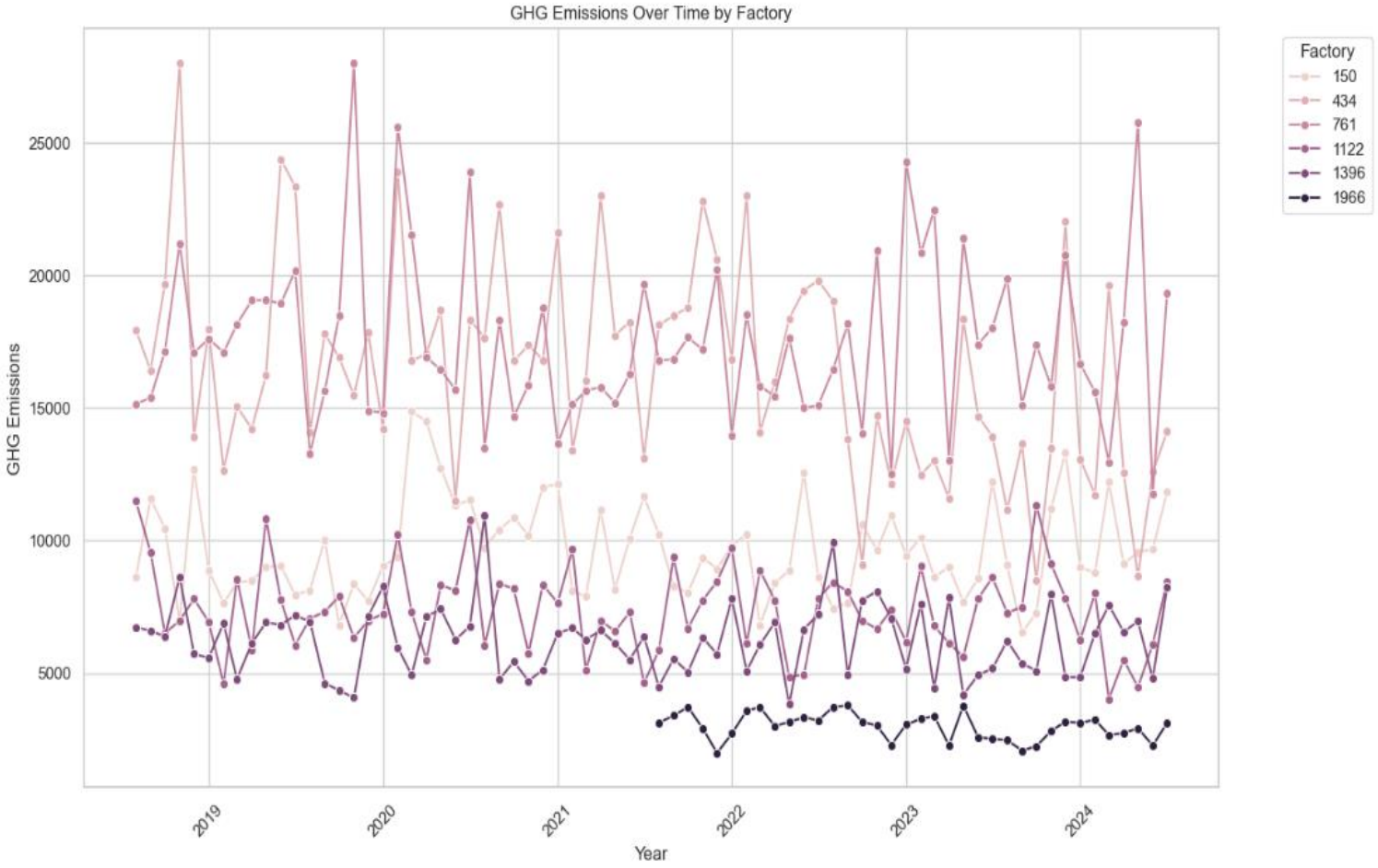


Figure 6 Yearly GHG Emission by Factories

Key Observations:

1. **Variation Across Factories:**
    - **Factory 434** consistently shows the highest GHG emissions, often peaking around 25,000 to 30,000 units.
    - **Factory 1966** shows the lowest emissions, with values mostly staying below 5,000 units.
2. **Consistency Over Time:**
    - Factories like **1966** and **150** have more stable emission levels over time, with fewer sharp peaks.
    - On the other hand, factories like **761** and **434** exhibit more volatility, with frequent and substantial fluctuations.

3. **Trends and Patterns:**
    - For most factories, there isn't a clear downward trend in emissions, indicating that there hasn't been a consistent effort to reduce emissions over this period.
    - Some factories, such as **761** and **434**, show sporadic spikes in emissions, which might be linked to specific periods of high production or inefficiencies.

4. **Recent Performance:**

As of 2023 and 2024, some factories, particularly **761** and **150**, show a slight upward trend, which could be concerning if the goal is to reduce GHG emissions by 10% by 2026.

4.1. GHG Emission and Efficiency Trends Over the Year by Factory

Key Observation:

- **Factory 150** shows a stable efficiency trend over the years with certain peaks.
- **Factory 434** and **Factory 1122** shows the least efficient results over the year leading to more GHG emission.
- **Factory 761** shows moderate to high efficiency. After 2022, it shows the unstable trend in the efficiency.
- After 2022, there is a significant change in trends in all the factories which should be looked upon.

4.2. GHG Emissions and Standard Allowed Hours Trend Over the Year by Factory

- SAH shows a direct relation with GHG Emissions, increasing SAH leads to high GHG Emissions
- Factory 150 shows a minimalistic change in GHG emission by increasing SAH, the reason could be they have high efficiency.
- During the recent years 2023 and 2024, there is a significant increase in SAH in Factory 761 leading to high GHG Emissions.
- After the year 2023, Factory 434 shows a decrease in SAH, leading to a decrease in GHG Emission.

4.3. GHG Emissions and Worker Trend Over the Year by Factory

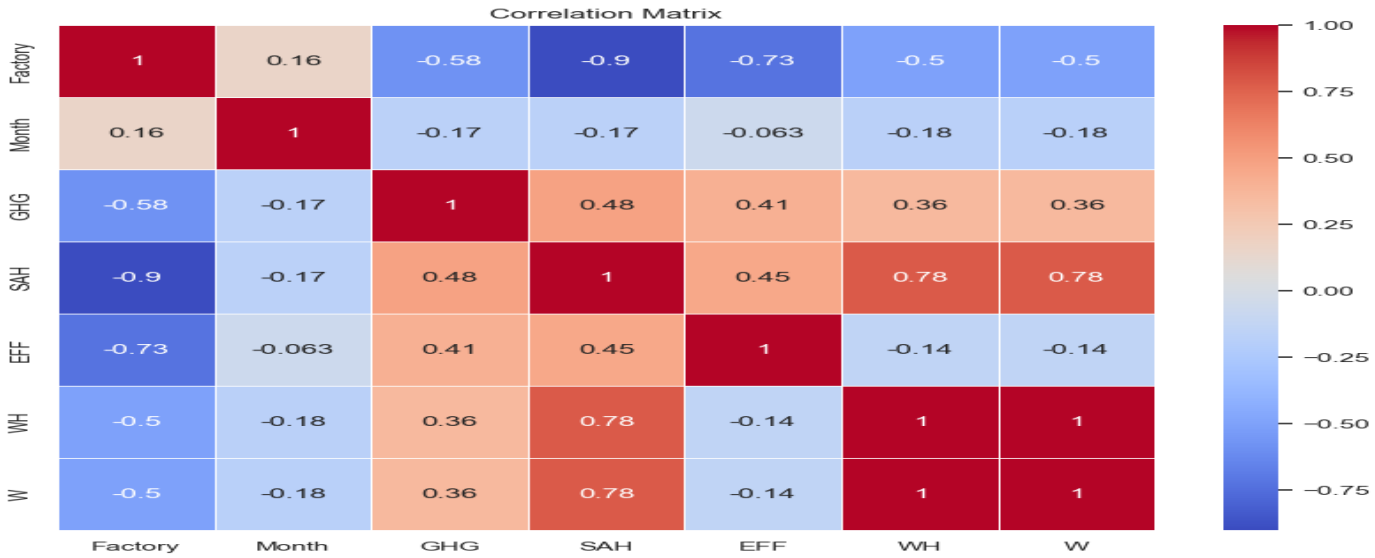
- Factory 761 shows a direct relation of GHG Emissions with workers.
- Factory 1966 shows a minimalistic no. of workers over the year but tend to have high GHG Emission.
- After 2023, there is a drastic increase in number of workers in Factory 434 leading to high GHG Emission, which can be due to high workload or high production units.

4.4. Standard Allowed Hours and Worker Trend Over the Year by Factory

- There is a drastic increase in Standard Allowed Hours with increase in workers after 2023 in Factory 761.
- Factory 150 performs well with moderate workers and moderate standard allowed hours.
- Factory 1966 shows high standard allowed hours with minimalistic number of workers which should be looked upon.

5. Correlation Matrix

The correlation matrix you've provided shows the relationships between different variables in the dataset. Correlation values range from -1 to 1.



Key Observations

1. GHG and Other Variables

- **GHG vs. SAH (0.48):** There is a moderate positive correlation between GHG emissions and Standard Allowed Hours (SAH). This suggests that as the time allowed for production increases, GHG emissions also tend to increase. This makes sense because more production time likely correlates with more energy consumption.
- **GHG vs. EFF (0.41):** There is a moderate positive correlation between GHG emissions and efficiency. Higher efficiency might be linked to higher production levels, leading to increased energy use and therefore higher emissions.
- **GHG vs. WH (0.36) and GHG vs. W (0.36):** Both Working Hours and the Number of Workers are moderately positively correlated with GHG emissions. More workers and more working hours likely result in more energy consumption, contributing to higher GHG emissions.

2. SAH and Other Variables

- **SAH vs. WH (0.78) and SAH vs. W (0.78):** There is a strong positive correlation between Standard Allowed Hours and both Working Hours and the Number of Workers. This suggests that more workers and longer working hours are required when the standard allowed time for production is higher.

3. EFF and Other Variables

- **EFF vs. SAH (0.45):** Efficiency and SAH have a moderate positive correlation. As efficiency improves, it might allow for more production within the same time, which might also explain the link with increased GHG emissions.
- **EFF vs. WH (-0.14) and EFF vs. W (-0.14):** There is a weak negative correlation between efficiency and both Working Hours and Number of Workers. This might indicate that higher efficiency can sometimes reduce the need for extended working hours and a larger workforce, though this relationship is not strong.

Model Training

In order to train the model, we first spit the data into features and target. Here ['SAH', 'EFF', 'WH', 'W'] are our features and ['GHG'] is our target. We drop the unnecessary columns for our model training like ['Factory', 'Month'].

After splitting the data, we train the data using different regression models. We use regression models because the data given is continuous and not discrete and evaluated their performance on the basis of Mean Square Error (MSE), R-Square value and Cross Validation Score.

- **Mean Squared Error (MSE)** is a metric used to measure the average squared difference between the actual values and the values predicted by a model. It provides a way to quantify the accuracy of a predictive model—the lower the MSE, the closer the predicted values are to the actual values.
- **R-squared**: This value indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.
- **Cross-Validation (CV) Score** is a measure used to evaluate the performance of a machine learning model by assessing how well it generalizes to an independent dataset. Cross-validation helps in understanding how the model will perform on unseen data and in preventing issues like overfitting.

Model Name	MSE	R2	CV Score
Linear Regressor	21201740.849	0.323264	19706809.8314
Decision Tree Regressor	24339290.2642	0.22311780	19530126.8020
Random Forest Regressor	7847981.233	0.74950145	10365542.6270
Extra Trees Regressor	7217825.651	0.76961529	9578253.1003
XG Boost	7537813.1773	0.75940165	12825296.39499

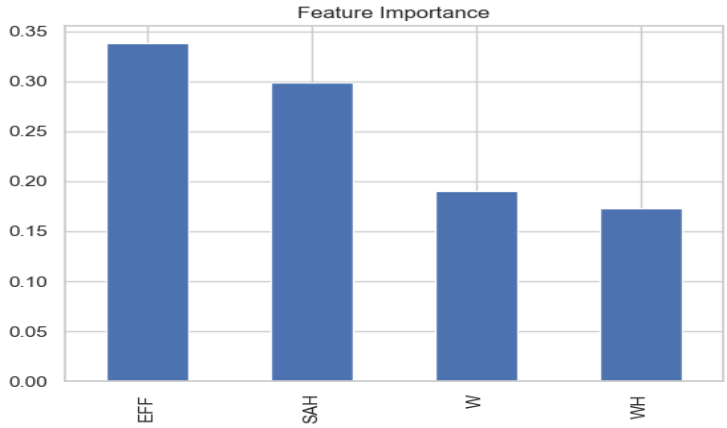
Here, Extra Trees Regressor performs relatively well with low MSE score as compared to other and approximately 77% of the variance in the target['GHG'] is explained by the model.

We further perform **Hyper Parameter Tuning.** Hyper parameter tuning is a crucial step in the machine learning pipeline that helps extract the full potential of models and ensures that they achieve the best possible performance on the task at hand. After Hyper tuning our model we get;

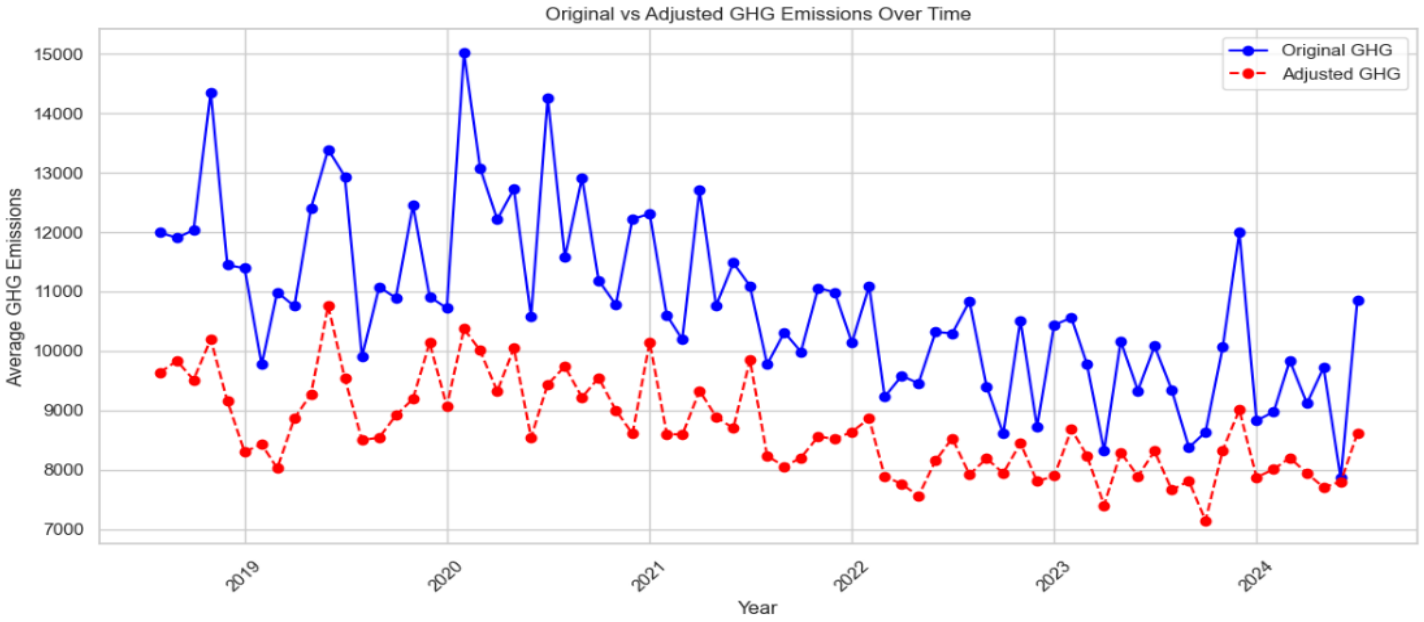
Extra Tress Regressor	*MSE: 6421151.622*	*R2: 0.79504*	*CV Score: 8268973.629*

Interpreting the GHG Emissions After Increasing Efficiency by 10%

The model Extra Trees Regressor also visualizes the feature importance. The most important feature was turn out to be Efficiency (EFF) followed by Standard Allowed Hours (SAH).



Let's see the change in Green House Gas Emission by increasing the efficiency by 10%, as this is the most important feature according to the model (Extra Trees Regressor) our data is trained.



After increasing the efficiency by 10%, we see a significant change in GHG emission as visualized in the graph. The average GHG reduction achieved by increasing efficiency is -8.70%, which align with our target of reducing 10% GHG by 2026.

Original GHG	Adjusted GHG
11843.4	12152.76
14111.3	14542.69
19339.7	7955.43
8470.5	5770.66
8236.1	7943.17

Recommendations for Reducing GHG Emissions:

Based on the analysis, several strategies can be implemented to effectively reduce greenhouse gas (GHG) emissions. The focus should be on targeting high-correlation factors, tailoring factory-specific strategies, and leveraging best practices from more efficient operations.

1. Target High-Impact Factors
Given the positive correlation between GHG emissions and variables such as Standard Allowed Hours (SAH), Efficiency (EFF), Working Hours (WH), and the Number of Workers (W), the following strategies are recommended:

- **Optimize Standard Allowed Hours (SAH):** Regularly review and adjust SAH to ensure they align with realistic production conditions, potentially reducing unnecessary energy consumption.

- **Improve Efficiency (EFF):** Focus on enhancing the efficiency of production processes to reduce time and resource consumption. This can be achieved through:
  - Standardizing best practices from the most efficient factories across all facilities.
  - Providing additional training to workers to improve skills and reduce production time.

- **Manage Workforce Size and Working Hours (W and WH):**
  - Streamline processes to ensure optimal use of the workforce and working hours, thereby minimizing energy use.
  - Implement energy-saving measures during non-production hours to reduce unnecessary energy consumption.

2. Factory-Specific Strategies
Tailored approaches are necessary for each factory, given the strong correlations between factory performance and variables like SAH and EFF:

- **Target High-Emission Factories: Factory 761 and Factory 434** should be prioritized as they contribute the most to overall GHG emissions. Focus on identifying and mitigating specific causes of high emissions in these factories, such as outdated equipment or inefficient processes.

- **Stabilize Volatile Emissions:** Factories with high volatility in emissions, such as **761 and 434**, should aim to stabilize their operations. Implementing better production planning and maintenance schedules could reduce these emission spikes and contribute to more consistent, lower emissions.

3. Leverage Best Practices
- **Learn from Low-Emission Factories:** Factories like **1966 and 150**, which have more consistent and lower emissions, should be studied to identify best practices. These practices should be implemented across other factories to drive overall improvements in sustainability.

4. Monitor and Adjust Continuously
- **Ongoing Monitoring:** Given the trend of increasing emissions in some factories, it is crucial to continuously monitor these trends and take corrective actions as needed. This will ensure that the factories stay on track to meet the target of reducing GHG emissions by 10% by 2026.